Solution

- Save raw data to AWS S3

- Remove incorrect, corrupted, incorrectly formatted, incomplete data and make it clean

- Check null values, count it, and replace it with N/A, check duplicates records and drop it

- Datasets which should be cleaned are Patients, Subscriber, Claims, Group\_subgroup

- Upload cleaned data to AWS Redshift

- Then create schema for all tables in redshift

- We need separate redshift table for use case output in a redshift schema

Use Cases

* Which disease has a maximum number of claims.
* Find those Subscribers having age less than 30 and they subscribe any subgroup
* Find out which group has maximum subgroups.
* Find out hospital which serve most number of patients
* Find out which subgroups subscribe most number of times
* Find out total number of claims which were rejected
* From where most claims are coming (city)
* Which groups of policies subscriber subscribe mostly Government or private
* Average monthly premium subscriber pay to insurance company.
* Find out Which group is most profitable
* List all the patients below age of 18 who admit for cancer
* List patients who have cashless insurance and have total charges greater than or equal for Rs. 50,000.
* List female patients over the age of 40 that have undergone knee surgery in the past year

Database Design

Tables Metadata Info

Claims Table

claim\_id, patient\_id, disease\_name, SUB\_ID, Claim\_Or\_Rejected, claim\_type, claim\_type, claim\_amount, claim\_date

Disease Table

SubGrpID, Disease\_ID, Disease\_name

Group Table

Country, premium\_written, zipcode, Grp\_Id, Grp\_Name, Grp\_Type, city, year

Hospital Table

Hospital\_id, hospital\_name, city, state, country

Patient\_records Table

Patient\_id, Patient\_name, patient\_gender, patient\_birth\_date, patient\_phone, disease\_name, city, hospital\_id

Subgroup Table

SubGrp\_id, SubGrp\_Name, Monthly\_Premium

Subscriber Table

sub\_id, first\_name, last\_name, Street, Birth\_date, Gender, Phone, Country, City, Zip code, Subgrp\_id, Elig\_ind, eff\_date, term\_date

Technologies and Platforms to be used in this solution

* AWS S3
* AWS Redshift
* Databricks
* AWS EMR Studio
* Pyspark
* Jira
* GitHub